Collocations

VL Sprachverarbeitung

Nils Reiter

April 14, 2022

Studienleistung: Glossareinträge

Metadaten (Definition)

Metadaten (english: Meta data) sind Daten über Daten. Metadaten eines sprachwissenschaftlichen Korpus enthalten Daten über das Korpus. Hierzu können folgende Informationen gehören:

die Sprache(n), in der die Daten vorliegen
Die Art der Daten (ist es gesprochene Sprache oder geschriebene Sprache?)
Wann wurden die Daten erhoben?
Wo wurden die Daten erhoben? (Gibt es ggf. regionale Phänomene wie Dialekte?)
Von wem wurden die Daten erhoben (Welche Forscher*innen, welches Institut/Uni)
Wo wurden die Daten veröffentlicht? (seriöse Fachzeitung oder ein privater Blog?)
Wer hat die Daten gespendet? (Muttersprachler*innen oder Nichtmuttersprachler*innen)

Bei gesprochenen Daten:
Auf welche Sound- und/oder Videodatei bezieht sich eine Transskription?
Wer spricht mit wem?
Gibt es eine Rollenverteilung?
Liegt eine Sprachstörung bei einem*einer oder mehreren Sprecher*innen vor?
Wann wurde welcher Token gesprochen?
Ist es eine natürliche Sprechsituation oder wurden die Daten unter Beobachtung erhoben?

Diese Daten müssen in bestimmten Dateiformaten gespeichert werden: XML, CSV, JSON etc.
Dabei spielt auch die Codierung der Daten eine Rolle. Bei Textdaten bietet sich UTF-8. Wenn ein Korpus in einem anderen Format codiert wurde, ist es sinnvoll, das Korpus in UTF-8 zu konvertieren.

Studienleistung: Glossareinträge

Type-Token-Relation

Die Type-Token-Relation, auch kurz TTR genannt, dient zur Unterscheidung zwischen einzelnen sprachlichen Äußerungen (**Token**) und der Klasse, der diesen Äußerungen zugrundeliegenden abstrakten Einheiten **(Types)**. Die TTR kann demnach als ein Instrument zur Messung der Wortschatzvielfalt verstanden werden und bildet den Quotienten aus Types und Token.

Beispiel: *Ein Text besteht aus 1.000 Wörtern und somit aus 1.000 Token. In diesem Text können Wörter öfter vorkommen, so dass es nur 400 unterschiedliche Wörter gibt, also 400 Types. Die Relation zwischen Types und Token würde in diesem Beispiel bei 40% liegen.*

Studienleistung: Glossareinträge

Types und Tokens (Definition)

Nicht alle Tokens sind Wörter (**Tokenisierung**). Types sind einzigartige Token. Sie werden innerhalb eines Textes nur einmal gezählt.
Beispiel: the cat chases the mouse
Token: the, cat, chases, the, mouse
Type: the, cat, chases, mouse

Section 1

Introduction

Introduction

A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things. *(MS99, p. 151)*

Introduction

A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things. *(MS99, p. 151)*

Examples

- »Das ist mein zweites Frühstück« (adjective noun)
- »Da müssen wir Abhilfe schaffen« (noun verb)
- »Es regnet in Strömen« (verb preposition noun)

Limited Compositionality

- Compositionality: The meaning of linguistic expressions can be understood from understanding their parts
- Collocations: Not entirely true
  - I.e., they are learned by heart and stored in lexicon

Limited Compositionality

- Compositionality: The meaning of linguistic expressions can be understood from understanding their parts
- Collocations: Not entirely true
  - I.e., they are learned by heart and stored in lexicon
- Related concepts
  - Idiomatic expressions, metaphors, figure of speech ...

Why are Collocations Interesting?

- Generation: Produce natural sounding expressions
  E.g., »Da müssen wir Abhilfe schaffen« instead of »Da müssen wir Abhilfe erzeugen«
- Parsing: Collocations are more likely to also be syntactic phrases
- Lexicography: Collocations should be included in dictionaries
- Social justice: Collocations may be important in reinforcing cultural stereotypes

How to Treat Collocations Quantitatively?

Multiple methods

- Frequency
- Mutual Information
- Mean and variance

Section 2

Frequency

Counting Bigrams

- Simple idea: We count bigrams (i.e., pairs of subsequent tokens)

Counting Bigrams

- Simple idea: We count bigrams (i.e., pairs of subsequent tokens)
- Corpus: Wikipedia pages (first $10\,000$ sentences)

Bigram	Frequency
wurde er	630
in der	623
wurde die	501
an der	386
mit dem	363
in die	362
in den	329
mit der	312
wurde das	291
wurde der	291
für die	248
er in	193
war er	181
von der	174
wo er	169
bei den	168
bei der	166
und wurde	165
an die	161
und die	150
er die	143
er als	142
er mit	142
wurden die	142
auf dem	135
für den	133
wurde sie	127
er zum	123
auf der	122

Counting Bigrams

- Simple idea: We count bigrams (i.e., pairs of subsequent tokens)
- Corpus: Wikipedia pages (first $10\,000$ sentences)
- Again, there are a lot of function words. Why?

Bigram	Frequency
wurde er	630
in der	623
wurde die	501
an der	386
mit dem	363
in die	362
in den	329
mit der	312
wurde das	291
wurde der	291
für die	248
er in	193
war er	181
von der	174
wo er	169
bei den	168
bei der	166
und wurde	165
an die	161
und die	150
er die	143
er als	142
er mit	142
wurden die	142
auf dem	135
für den	133
wurde sie	127
er zum	123
auf der	122

Counting Bigrams

- Simple idea: We count bigrams (i.e., pairs of subsequent tokens)
- Corpus: Wikipedia pages (first $10\,000$ sentences)
- Again, there are a lot of function words. Why?
- Zipf's law: Two words that are highly frequent have much higher chance to co-occur with high frequency

Bigram	Frequency
wurde er	630
in der	623
wurde die	501
an der	386
mit dem	363
in die	362
in den	329
mit der	312
wurde das	291
wurde der	291
für die	248
er in	193
war er	181
von der	174
wo er	169
bei den	168
bei der	166
und wurde	165
an die	161
und die	150
er die	143
er als	142
er mit	142
wurden die	142
auf dem	135
für den	133
wurde sie	127
er zum	123
auf der	122

Counting Bigrams

Content Words

- Content words: Nouns, verbs, adjectives, adverb
- My operationalization here: Remove everything that doesn't contain one upper-case letter
  - Because verb-verb combinations are rare (as bigrams)
  - But we're missing verb-adverb combinations

Counting Bigrams

Content Words

- Content words: Nouns, verbs, adjectives, adverb
- My operationalization here: Remove everything that doesn't contain one upper-case letter
  - Because verb-verb combinations are rare (as bigrams)
  - But we're missing verb-adverb combinations

Bigram	Frequency
Jahre alt	56
Bevölkerung waren	47
Prozent waren	46
Jahre später	45
of Fame	44
Hall of	43
New York	41
als Nachfolger	41
Olympischen Spielen	35
Professor für	32
ersten Mal	32
er Mitglied	29
Fame aufgenommen	28
selben Jahr	28
Zweiten Weltkrieg	26
zum Mitglied	25
zum Professor	24
Jahr später	23
zwei Jahre	22
University of	21
Professor an	20
nach Deutschland	20
Betrieb genommen	18
Bevölkerung war	18
Los Angeles	18
drei Jahre	18
als Professor	17
Im Jahr	16
Lehrstuhl für	16

Focus Words

- Look at bigrams that contain a specific word
- In this case: »Gründen«

Focus Words

- Look at bigrams that contain a specific word
- In this case: »Gründen«

Bigram	Frequency
gesundheitlichen Gründen	7
Gründen von	3
finanziellen Gründen	2
Gründen abgeben	1
Gründen als	1
Gründen auf	1
Gründen aus	1
Gründen den	1
Gründen die	1
Gründen gab	1
Gründen ihre	1
Gründen interessierte	1
Gründen nach	1
Gründen um	1
Gründen zurück	1
disziplinarischen Gründen	1
gesundheitlichen Problemen	1
nationalpolitischen Gründen	1
paläographischen Gründen	1
persönlichen Gründen	1
politischen Gründen	1
strategischen Gründen	1

Section 3

Point-wise Mutual Information

Introduction

Example

»1910 wurde Gerland _____ _____ in Jena.«

- Knowing one word makes predicting the next easier
- One word provides information about the next – it reduces insecurity

Introduction

Example

»1910 wurde Gerland außerordentlicher _____ in Jena.«

- Knowing one word makes predicting the next easier
- One word provides information about the next – it reduces insecurity

Introduction

Example

»1910 wurde Gerland außerordentlicher Professor in Jena.«

- Knowing one word makes predicting the next easier
- One word provides information about the next – it reduces insecurity

Intuition

- We are interested in the (potential) collocation »außerordentlicher Professor«

Intuition

Word	Counts	Frequency
außerordentlicher	109	5.5×10^{-6}
Professor	2126	1×10^{-4}
All	19 811 129	1

- We are interested in the (potential) collocation »außerordentlicher Professor«
- We interpret counts as probabilities
  - If we pick a random word, the probability that it is »Professor«, is 1×10^{-4}: $p(\text{Professor}) = 1 \times 10^{-4} \simeq 0.000\,107\,31$

Intuition

Word	Counts	Frequency
außerordentlicher	109	5.5×10^{-6}
Professor	2126	1×10^{-4}
All	19 811 129	1

- We are interested in the (potential) collocation »außerordentlicher Professor«
- We interpret counts as probabilities
  - If we pick a random word, the probability that it is »Professor«, is 1×10^{-4}: $p(\text{Professor}) = 1 \times 10^{-4} \simeq 0.000\,107\,31$
  - If we pick two random words, how likely is it that they are »außerordentlicher«, followed by »Professor«?

Intuition

Word	Counts	Frequency
außerordentlicher	109	5.5×10^{-6}
Professor	2126	1×10^{-4}
All	19 811 129	1

- We are interested in the (potential) collocation »außerordentlicher Professor«
- We interpret counts as probabilities
  - If we pick a random word, the probability that it is »Professor«, is 1×10^{-4}: $p(\text{Professor}) = 1 \times 10^{-4} \simeq 0.000\,107\,31$
  - If we pick two random words, how likely is it that they are »außerordentlicher«, followed by »Professor«? $p(\text{außerordentlicher}) \times p(\text{Professor}) = 5.5 \times 10^{-10}$

Intuition

Word	Counts	Frequency
außerordentlicher	109	5.5×10^{-6}
Professor	2126	1×10^{-4}
All	19 811 129	1

- We are interested in the (potential) collocation »außerordentlicher Professor«
- We interpret counts as probabilities
  - If we pick a random word, the probability that it is »Professor«, is 1×10^{-4}: $p(\text{Professor}) = 1 \times 10^{-4} \simeq 0.000\,107\,31$
  - If we pick two random words, how likely is it that they are »außerordentlicher«, followed by »Professor«? $p(\text{außerordentlicher}) \times p(\text{Professor}) = 5.5 \times 10^{-10}$
- This is the probability that these two words appear together – if they are distributed randomly
  - Denominator (Nenner) of the point-wise mutual information!
  - The ›real‹ probability is the numerator (Zähler)

Independence of Events and Probabilities

Interpretations

$$I(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

- Fraction between real and expected co-occurrence probability

Interpretations

$$I(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

- Fraction between real and expected co-occurrence probability
- Thought experiments
  - No dependence – co-occurrence has same probability as by chance
    - $p(w_1) = 0.01, p(w_2) = 0.01, p(w_1, w_2) = 0.0001, \Rightarrow I(w_1, w_2) = \log_2 1 = 0$

Interpretations

$$I(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

- Fraction between real and expected co-occurrence probability
- Thought experiments
  - No dependence – co-occurrence has same probability as by chance
    - $p(w_1) = 0.01, p(w_2) = 0.01, p(w_1, w_2) = 0.0001, \Rightarrow I(w_1, w_2) = \log_2 1 = 0$
  - Co-occurrence is 8 times more probable than by chance
    - $p(w_1) = 0.01, p(wR_2) = 0.01, p(w_1, w_2) = 0.008, \Rightarrow I(w_1, w_2) = \log_2 \frac{0.0008}{0.0001} = 3$

Interpretations

$$I(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

- Fraction between real and expected co-occurrence probability
- Thought experiments
  - No dependence – co-occurrence has same probability as by chance
    - $p(w_1) = 0.01, p(w_2) = 0.01, p(w_1, w_2) = 0.0001, \Rightarrow I(w_1, w_2) = \log_2 1 = 0$
  - Co-occurrence is 8 times more probable than by chance
    - $p(w_1) = 0.01, p(wR_2) = 0.01, p(w_1, w_2) = 0.008, \Rightarrow I(w_1, w_2) = \log_2 \frac{0.0008}{0.0001} = 3$
  - Co-occurrence is 8 times less probable than by chance
    - $p(w_1) = 0.01, p(w_2) = 0.01, p(w_1, w_2) = 0.0000125, \Rightarrow I(w_1, w_2) = \log_2 \frac{0.0000125}{0.0001} = -3$

Point-wise Mutual Information

MS99, pp. 178 ff.

- Point-wise: Statement about values of random variable (i.e., occurrence of specific word)
  - Non-pointwise mutual information makes a statement about random variables themselves

Point-wise Mutual Information

MS99, pp. 178 ff.

- Point-wise: Statement about values of random variable (i.e., occurrence of specific word)
  - Non-pointwise mutual information makes a statement about random variables themselves
- Mutual: Symmetric
  - One word provides information to the next and vice versa

Point-wise Mutual Information

MS99, pp. 178 ff.

- Point-wise: Statement about values of random variable (i.e., occurrence of specific word)
  - Non-pointwise mutual information makes a statement about random variables themselves
- Mutual: Symmetric
  - One word provides information to the next and vice versa

$$I(w_1, w_2) \quad = \quad \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

Point-wise Mutual Information

MS99, pp. 178 ff.

- Point-wise: Statement about values of random variable (i.e., occurrence of specific word)
  - Non-pointwise mutual information makes a statement about random variables themselves
- Mutual: Symmetric
  - One word provides information to the next and vice versa

$$
\begin{aligned}
I(w_1, w_2) &= \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \\
p(w_i) &= \text{Probability of word } w_i \\
p(w_i, w_j) &= \text{Probability of both words appearing together, up to a certain distance}
\end{aligned}
$$

Point-wise Mutual Information

MS99, pp. 178 ff.

- Point-wise: Statement about values of random variable (i.e., occurrence of specific word)
  - Non-pointwise mutual information makes a statement about random variables themselves
- Mutual: Symmetric
  - One word provides information to the next and vice versa

$$
\begin{aligned}
I(w_1, w_2) &= \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \\
p(w_i) &= \text{Probability of word } w_i \\
p(w_i, w_j) &= \text{Probability of both words appearing together, up to a certain distance} \\
\log_2 x = y &\equiv 2^y = x
\end{aligned}
$$

Section 4

Mean and Variance

Assumption in Bigram Counting

- Fixed word order
  - »gesundheitliche« comes directly before »Gründe«
- But: Modified collocations are still recognizable
  - »gesundheitliche vorgeschobene Gründe«
- Increasing n for n-grams doesn't help
  - Because it decreases the frequencies

Assumption in Mutual Information

- Fixed window size
  - Define a window size w (e.g., 5 words)
  - Count frequency with which two words appear $\leq w$ words apart

Assumption in Mutual Information

- Fixed window size
  - Define a window size w (e.g., 5 words)
  - Count frequency with which two words appear $\leq w$ words apart
- Regularly used in NLP
- How to select window size?
  - Empirical: See what works best
- Intuitions
  - Small windows emphasize syntactic properties
  - Large windows emphasize semantic properties

Third Option: Co-Dispersion / Mean and Variance

- Look at mean and variance of distances between two words

Third Option: Co-Dispersion / Mean and Variance

- Look at mean and variance of distances between two words

Example

Am	15.	November	1734	legte	er	aus	gesundheitlichen	Gründen	die	Professur	nieder	.

Third Option: Co-Dispersion / Mean and Variance

- Look at mean and variance of distances between two words

Example

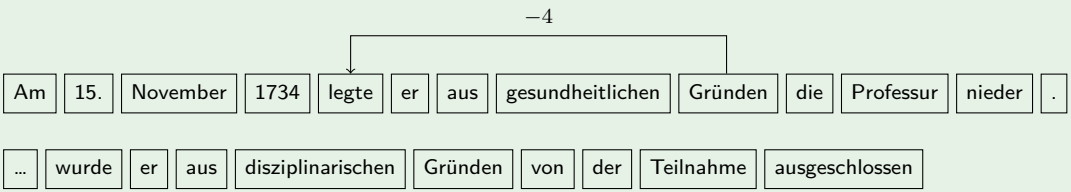
Am | 15. | November | 1734 | legte | er | aus | gesundheitlichen | Gründen | die | Professur | nieder | .

... | wurde | er | aus | disziplinarischen | Gründen | von | der | Teilnahme | ausgeschlossen

Third Option: Co-Dispersion / Mean and Variance

- Look at mean and variance of distances between two words

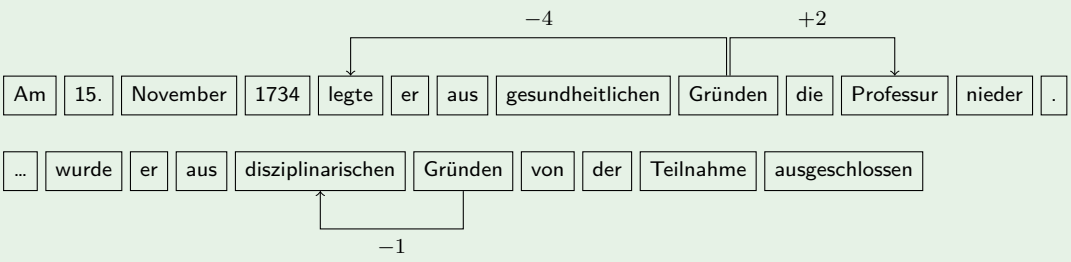
Example



Third Option: Co-Dispersion / Mean and Variance

- Look at mean and variance of distances between two words

Example



Mean and Variance

- Mean: Arithmetic mean, average

$$\mu = \frac{1}{N} \sum_{i=1}^{N} i$$

Mean and Variance

- Mean: Arithmetic mean, average

$$\mu = \frac{1}{N}\sum_{i=1}^{N} i$$

- Variance
  - Measures the dispersion around the man
  - What is the average distance between the data points and the mean?

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(i-\mu)^2$$

  - Derived notion: Standard deviation

$$\sigma = \sqrt{\sigma^2}$$

Mean and Variance

- Mean: Arithmetic mean, average

$$\mu = \frac{1}{N}\sum_{i=1}^{N} i$$

- Variance
  - Measures the dispersion around the man
  - What is the average distance between the data points and the mean?

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(i-\mu)^2$$

  - Derived notion: Standard deviation

$$\sigma = \sqrt{\sigma^2}$$

- What does this tell us?

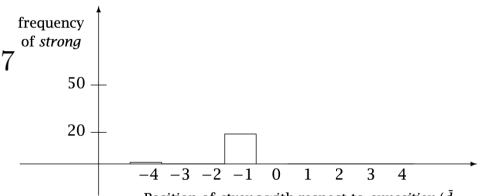
Example: »strong support« vs. »strong opposition«

MS99, p. 160

Example: »strong support« vs. »strong opposition«

MS99, p. 160

- »strong opposition«: $\mu = -1.15; \sigma = 0.67$

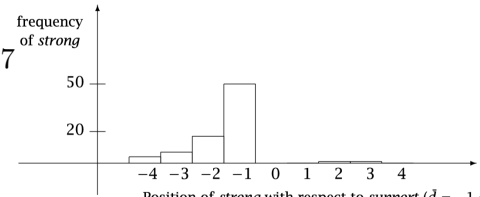


Position of *strong* with respect to *opposition* ($\bar{d} = -1.15, s = 0.67$).

Example: »strong support« vs. »strong opposition«

MS99, p. 160

- »strong opposition«: $\mu = -1.15; \sigma = 0.67$
- »strong support«: $\mu = -1.45; \sigma = 1.07$

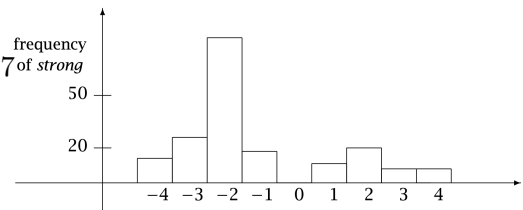


Position of *strong* with respect to *support* ($\bar{d} = -1.45, s = 1.07$).

Example: »strong support« vs. »strong opposition«

MS99, p. 160

- »strong opposition«: $\mu = -1.15; \sigma = 0.67$
- »strong support«: $\mu = -1.45; \sigma = 1.07$
- »strong for«: $\mu = -1.12; \sigma = 2.05$



Position of *strong* with respect to *for* ($\bar{d} = -1.12, s = 2.15$).

Example: »strong support« vs. »strong opposition«

MS99, p. 160

- »strong opposition«: $\mu = -1.15; \sigma = 0.67$
- »strong support«: $\mu = -1.45; \sigma = 1.07$
- »strong for«: $\mu = -1.12; \sigma = 2.05$
- Low deviation: Words appear at the same position
- The lower, the more stable is this collocation

Section 5

Summary

Summary

Collocations

- Words that ›often‹ appear together
- Point-wise mutual information
- Counting bigrams
- Co-Dispersion: Look at mean distances and their dispersion